

**ADVANCED PROTEOMIC CHARACTERIZATION OF THE 26S PROTEASOME IN  
ARABIDOPSIS REVEALS INSIGHTS INTO COMPOSITION AND ASSEMBLY**

by

David C. Gemperline

A dissertation submitted in partial fulfillment of  
the requirements for the degree of

Doctor of Philosophy

(Genetics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2016

Date of final oral examination: TBD

The dissertation is approved by the following members of the Final Oral Committee:

Richard D. Vierstra, Professor, Genetics

Richard Amasino, Professor, Genetics

Josh Coon, Professor, Chemistry

Donna Fernandez, Professor, Botany

Patrick Masson, Professor, Genetics

© Copyright by David C. Gemperline 2016

All Rights Reserved

*To Erin, my wife.*

## ACKNOWLEDGMENTS

*Science doesn't purvey absolute truth. Science is a mechanism, a way of trying to improve your knowledge of nature. It's a system for testing your thoughts against the universe and seeing whether they match. This works not just for the ordinary aspects of science, but for all of life.*

— ISAAC ASIMOV (1988)

Acknowledgements go here.

## TABLE OF CONTENTS

Table of Contents	iii
List of Tables	iv
List of Figures	v
List of Abbreviations and Acronyms	vi
Abstract	vii
Abstract	ix
Chapter 1: The Ubiquitin 26S Proteasome System	1
1.1 Ubiquitin Conjugating Machinery . . . . .	1
1.1.1 E1s . . . . .	1
1.1.2 E2s . . . . .	1
1.1.3 E3s . . . . .	1
1.1.4 DUBS . . . . .	1
1.2 The 26S Proteasome . . . . .	1
1.2.1 The 20S Core Protease . . . . .	1
1.2.2 The 19S Regulatory Particle . . . . .	1

1.3	Proteasome Expression . . . . .	2
1.4	Proteasome Assembly . . . . .	2
1.5	Proteasome Post-Translational Modification . . . . .	2
1.6	Proteasome Degredation . . . . .	2
1.7	Proteasome Interacting Proteins . . . . .	2
Chapter 2: Morpheus Spectral Counter: A Computational Tool for Label-		
Free Quantitative Mass Spectrometry using the Morpheus Search		
	Engine	3
2.1	Summary . . . . .	3
2.2	Main Text . . . . .	4
2.3	Methods . . . . .	17
2.4	Tutorial . . . . .	17
2.5	Acknowledgements . . . . .	17
2.6	References . . . . .	18
Colophon		22

## LIST OF TABLES

2.1	Table of adj_NSAF values for each 20S proteasome subunit (equivalent to dNSAF) generated by analyzing the proteasome spike in experiments with the TPP and ABACUS pipeline . . . . .	13
-----	--	----

## LIST OF FIGURES

2.1	MSpC Graphic User Interface (GUI) and software flow chart . . . . .	6
2.2	Confirmation of MSpC accuracy by analysis of MS/MS datasets generated with the Universal Proteome Standard 2 (UPS2) . . . . .	8
2.3	Re- Analysis of MS/MS datasets generated with the Universal Proteome Standard 2 (UPS2) . . . . .	10
2.4	Confirmation of MSpC accuracy by analysis of MS/MS datasets generated with affinity purified Arabidopsis 20S proteasomes spiked into a total cell lysate from E. coli. Following MS/MS analysis, the dNSAF and uNSAF values for each subunit/isoform were determined by Morpheus and MSpC . . . . .	14



## LIST OF ABBREVIATIONS AND ACRONYMS

API	Application programming interface
BCA	Bicinchoninic acid protein assay
BLAST	Basic local alignment search tool
C#	C sharp, a programming language
Da	Dalton, the atomic mass unit
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
ESI	Electrospray ionization
E-value	Expectation value
FASTA	A format for storing protein sequences
FDR	False discovery rate
GUI	Graphical user interface
HCD	Higher-energy collisional dissociation
HPLC	High-performance liquid chromatography
LC	Liquid chromatography
<i>m</i>	Mass
min	Minute
MS	Mass spectrometry

MS <sup>1</sup>	Survey mass analysis
MS/MS	Tandem mass spectrometry
NCE	Normalized collision energy
nLC	Nanoflow liquid chromatography
ppm	Part per million
PSM	Peptide-spectrum match
PTM	Post-translational modification
s	Second
SILAC	Stable isotope labeling by amino acids in cell culture
S/N	Signal-to-noise ratio
TMT	Tandem mass tag

**ADVANCED PROTEOMIC CHARACTERIZATION OF THE 26S  
PROTEASOME IN ARABIDOPSIS REVEALS INSIGHTS INTO  
COMPOSITION AND ASSEMBLY**

David C. Gemperline

Under the supervision of Professor Richard D. Vierstra

At the University of Wisconsin-Madison

**Abstract**

The 26S proteasome is the central proteolytic effector in the ubiquitin system that is responsible for degrading numerous regulators following their selective ubiquitylation. While much is known about the construction of the yeast and mammalian particles, little is known about the pathways used to assemble the plant particle. One challenge is that the known yeast chaperones appear sufficiently diverged to preclude high-confidence identification of their plant counterparts by genomic searches. Here, we used in-depth mass spectrometric analysis of Arabidopsis 26S proteasomes, which were affinity purified from seedlings under conditions that promote the accumulation of assembly intermediates, to identify a large collection of interacting proteins that associate with either the core protease (CP) or regulatory particle (RP). Sequence comparisons, Y2H and BiFC studies revealed

that some are likely assembly chaperones, with several CP factors harboring the signature C-terminal HbYX motif that allows their association with the  $\alpha$ -subunit ring. Several of the RP-specific factors appear to be orthologs of the chaperones Nas2, Nas6, Hsm3 and Ecm29. Whereas yeast assembles only a single particle type, mammals can assemble alternate proteasomes by replacing individual subunits with distinct isoforms (e.g., immunoproteasomes). In plants, most 26S proteasome subunits are encoded by paralogous genes with sufficient divergence to suggest that plants also accumulate a collection of particles. However, proteomic analysis of proteasomes selectively enriched using paralog-specific tags strongly imply that although plants possess this genetic diversity, the incorporation of these paralogs appears random, and is mainly influenced by the differential expression of the corresponding genes. Taken together, these proteomic studies provide the first insights into plant proteasome assembly and diversity, and identify factors that build the CP and RP subcomplexes and finally the 26S holo-particle.

Richard D. Vierstra

## ABSTRACT

### Abstract

The 26S proteasome is the central proteolytic effector in the ubiquitin system that is responsible for degrading numerous regulators following their selective ubiquitylation. While much is known about the construction of the yeast and mammalian particles, little is known about the pathways used to assemble the plant particle. One challenge is that the known yeast chaperones appear sufficiently diverged to preclude high-confidence identification of their plant counterparts by genomic searches. Here, we used in-depth mass spectrometric analysis of Arabidopsis 26S proteasomes, which were affinity purified from seedlings under conditions that promote the accumulation of assembly intermediates, to identify a large collection of interacting proteins that associate with either the core protease (CP) or regulatory particle (RP). Sequence comparisons, Y2H and BiFC studies revealed that some are likely assembly chaperones, with several CP factors harboring the signature C-terminal HbYX motif that allows their association with the  $\alpha$ -subunit ring. Several of the RP-specific factors appear to be orthologs of the chaperones Nas2, Nas6, Hsm3 and Ecm29. Whereas yeast assembles only a single particle type, mammals can assemble alternate proteasomes by replacing individual subunits with distinct isoforms (e.g., immunoproteasomes). In plants, most 26S proteasome

subunits are encoded by paralogous genes with sufficient divergence to suggest that plants also accumulate a collection of particles. However, proteomic analysis of proteasomes selectively enriched using paralog-specific tags strongly imply that although plants possess this genetic diversity, the incorporation of these paralogs appears random, and is mainly influenced by the differential expression of the corresponding genes. Taken together, these proteomic studies provide the first insights into plant proteasome assembly and diversity, and identify factors that build the CP and RP subcomplexes and finally the 26S holo-particle.

## **Chapter 1**

### **THE UBIQUITIN 26S PROTEASOME SYSTEM**

#### **1.1. Ubiquitin Conjugating Machinery**

##### **1.1.1. E1s**

##### **1.1.2. E2s**

##### **1.1.3. E3s**

##### **1.1.4. DUBS**

#### **1.2. The 26S Proteasome**

##### **1.2.1. The 20S Core Protease**

##### **1.2.2. The 19S Regulatory Particle**

###### **1.2.2.1. Regulatory Particle Base**

###### **1.2.2.2. Regulatory Particle Lid**

### **1.3. Proteasome Expression**

### **1.4. Proteasome Assembly**

### **1.5. Proteasome Post-Translational Modification**

### **1.6. Proteasome Degredation**

### **1.7. Proteasome Interacting Proteins**



## Chapter 2

# **MORPHEUS SPECTRAL COUNTER: A COMPUTATIONAL TOOL FOR LABEL-FREE QUANTITATIVE MASS SPECTROMETRY USING THE MORPHEUS SEARCH ENGINE**

### 2.1. Summary

Label-free quantitative MS based on the Normalized Spectral Abundance Factor (NSAF) has emerged as a straightforward and robust method to determine the relative abundance of individual proteins within complex mixtures. Here, we present Morpheus Spectral Counter (MSpC) as the first computational tool that directly calculates NSAF values from output obtained from Morpheus, a fast, open-source, peptide-MS/MS matching engine compatible with high-resolution accurate-mass instruments. NSAF has distinct advantages over other MS-based quantification methods, including a higher dynamic range as compared to isobaric tags, no requirement to align and re-extract MS1 peaks, and increased speed. MSpC features an easy to use graphic user interface that additionally calculates both distributed and unique NSAF values to permit analyses of both protein families and isoforms/proteoforms. MSpC determinations of protein concentration were linear over

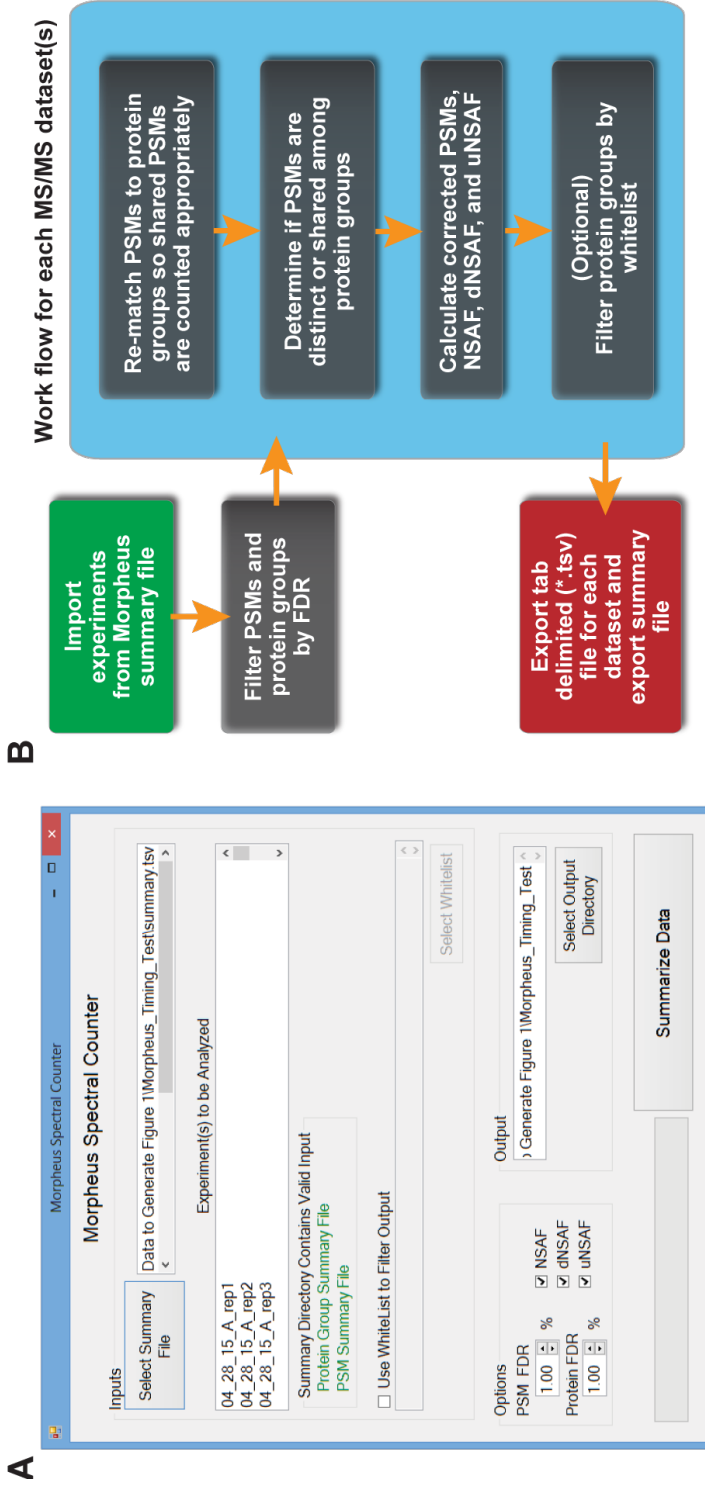
several orders of magnitude based on the analysis of several high-mass accuracy datasets either obtained from PRIDE or generated with total cell extracts spiked with purified Arabidopsis 20S proteasomes. The MSpC software was developed in C# and is open sourced under a permissive license with the code made available at [http://dcgemperline.github.io/Morpheus\\_SpC/](http://dcgemperline.github.io/Morpheus_SpC/).

## 2.2. Main Text

Quantification of individual polypeptides within complex mixtures by MS is an extremely useful tool to understand proteomic changes in organisms during growth and development, and after environmental perturbation (Wong and Cagney, 2010). While a number of MS/MS strategies have been developed to measure protein abundance, including Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC), labeling with isobaric tags, and Absolute Quantification of proteins (AQUA) (Gerber et al., 2003; Ong et al., 2002; Ross et al., 2004; Thompson et al., 2003), label-free quantification (LFQ) have become increasingly popular given their simplicity and low cost (Wong and Cagney, 2010; Zhang et al., 2006). One LFQ strategy infers abundance from the number of observed peptide spectra matches (PSMs). For these PSM-based approaches, changes in protein abundance can be generated artifactually when total PSMs differ among samples and because longer proteins

tend to produce more raw counts. For these reasons normalizing for both protein length and total PSMs is paramount. While this adjustment can be made in a number of ways; one of the most straight forward methods is to use Normalized Spectral Abundance Factor (NSAF), a length- and count-normalized measure for each protein (Zybailov et al., 2006). Further improvements to the NSAF algorithm have been made by accounting for shared peptides in distributed NSAF (dNSAF), which distributes common PSMs among a family of isoforms/proteoforms based on the number of distinct PSMs observed for each isoform/proteoform, and unique NSAF (uNSAF), which ignores shared PSMs and only assigns distinct PSMs to each specific isoform/proteoform (Zhang et al., 2010).

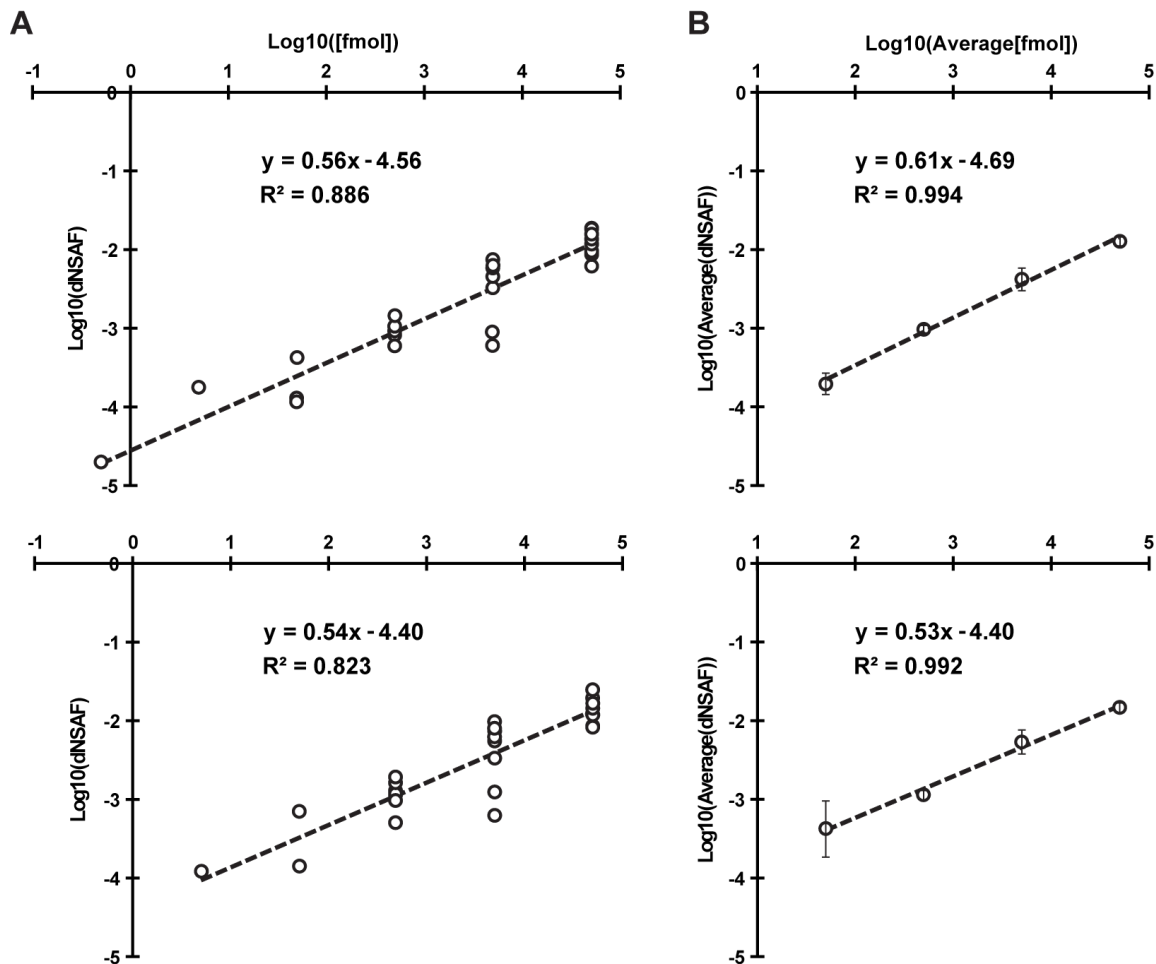
The Morpheus MS search engine was recently designed for high-resolution, accurate-mass data obtained from Orbitrap-based instruments to provide faster matching of spectra to peptides (Wenger and Coon, 2013). Unfortunately, no downstream automated tools are available to facilitate LFQ analysis, which can be quite challenging, if not impossible, to complete manually when accounting for shared peptides. To overcome this bottleneck, we developed Morpheus Spectral Counter (MSpC) as the first LFQ computational tool that integrates directly with Morpheus to calculate NSAF, dNSAF, uNSAF, and corrected PSM (Fermin et al., 2011) values in complex protein samples. MSpC is fully automated, and only requires a



**Figure 2.1: MSpc Graphic User Interface (GUI) and software flow chart.** (A) Screenshot of the GUI. The input requires the user to select a Morpheus search summary file containing experiments to be analyzed. The user can optionally select a whitelist to filter output, and select an output directory. Additional options can set peptide and protein FDR cutoffs, and method of quantification for output, including Normalized Spectral Abundance Factor (NSAF), distributed NSAF (dNSAF), and unique NSAF (uNSAF). A progress bar highlights completion of the analysis. (B) Data analysis flow chart. Experiments and groups of experiments to be analyzed are imported through the Morpheus summary file. PSMs and protein groups are filtered at the specified FDR cutoff with a default of 1%.

Morpheus search summary file (summary.tsv) as input. The user interface (see Figure 2.1 (A)) allows one to select the summary file and displays the raw MS/MS files that will be analyzed by MSpC. Due to shared peptides being attributed to only one instance of a protein group in Morpheus's PSM file, PSMs are re-matched to all possible protein groups. PSMs are then cataloged as shared or as unique (distinctly matching one protein group) to generate NSAF, dNSAF, and uNSAF outputs. Finally, the output can be filtered for proteins of interest by specifying a comma delimited file containing unique identifiers and descriptions. Some important features of MSpC are its ability to handle fractionation experiments as input, and the ability to whitelist proteins of interest in the output by specifying a csv file (see Tutorial). Options exist to specify global PSM and protein group FDR rates (thus avoiding increased FDRs when one analyzes many experiments at once), to output NSAF, dNSAF, and uNSAF values, to require a minimum number of unique peptides to quantify a protein, and to specify an output directory. A progress bar indicates completion of the analysis by MSpC. To validate the accuracy of MSpC, we analyzed two MS/MS datasets available in PRIDE that were previously generated by high-energy collision-induced dissociation using Thermo Q-Exactive Orbitrap instruments.

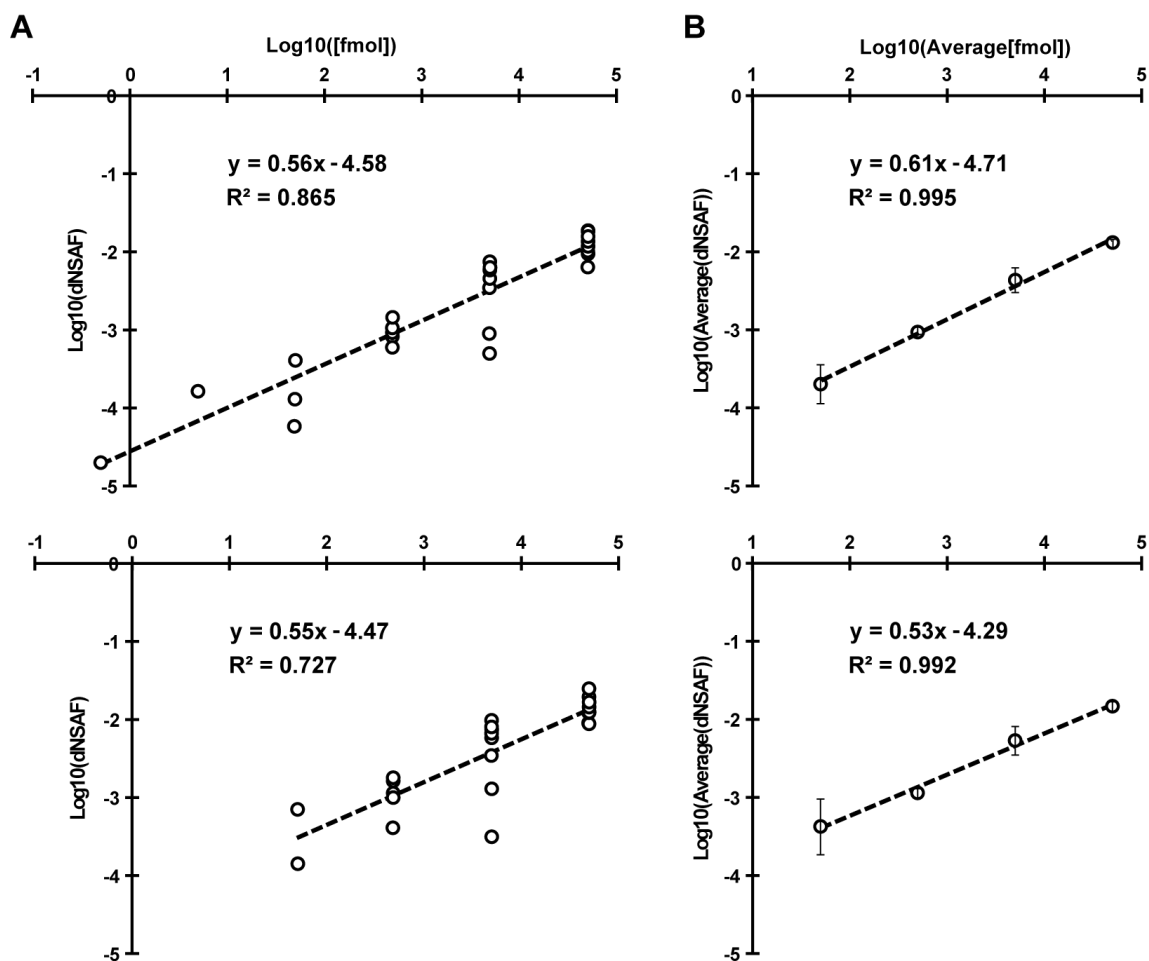
Here, *Xenopus* egg (see top, Figure 2.2) and embryo (bottom, Figure 2.2) extracts



**Figure 2.2: Confirmation of MSpC accuracy by analysis of MS/MS datasets generated with the Universal Proteome Standard 2 (UPS2).** The array of UPS2 standards were spiked into *Xenopus laevis* egg (Top) and embryo (Bottom) extracts at a range of concentrations. Following MS/MS analysis, dNSAF values for each protein were determined by Morpheus and MSpC. (A) A log-log plot of dNSAF versus concentration for each UPS2 protein detected across each fmol range. (B) A log-log plot of average dNSAF vs average concentration of each group of UPS2 proteins at each fmol range: (50, 500, 5000, and 50,000 fmol).

were spiked at a 4:1 ratio with the Universal Proteome Standard 2 (UPS2), a mix of 48 purified proteins at defined molar ratios of 0.5, 5, 50, 500, 5000, and 50,000, with each ratio containing a different set of 8 of the 48 proteins. As shown in Figure 2.2 A, when the Morpheus/MSpC pipeline was used to calculate the average dNSAF value for each UPS2 protein, requiring only a single unique peptide to quantify, strong linear correlations ( $R^2 = 0.886$  and  $0.823$ ) were obtained across a 1,000 fold change in abundance (50 fmol to 50,000 fmol). In fact, the  $R^2$  values were similar to those obtained by others with PSM-based LFQ methods (Cox et al., 2014; Tu et al., 2014). This linear correlation was further strengthened when the dNSAF values were averaged for all UPS2 proteins within each of the concentration groups, with  $R^2$  values of 0.994 and 0.992 for the egg and embryo datasets, respectively (see Figure 2.2 B). Notably, the slope of the concentration series was significantly less than unity, showing that NSAF measurements are not appropriate for absolute quantification, which was expected given that NSAF is a relative value.

We also reprocessed the UPS2 dataset using the option of requiring a minimum of two unique peptides for quantification, which should improve stringency. This option provided only a minor improvement in overall linearity for the average UPS2 dNSAF values, but decreased linearity when each UPS2 protein was considered individually and removed some UPS2 proteins at low concentrations (compare

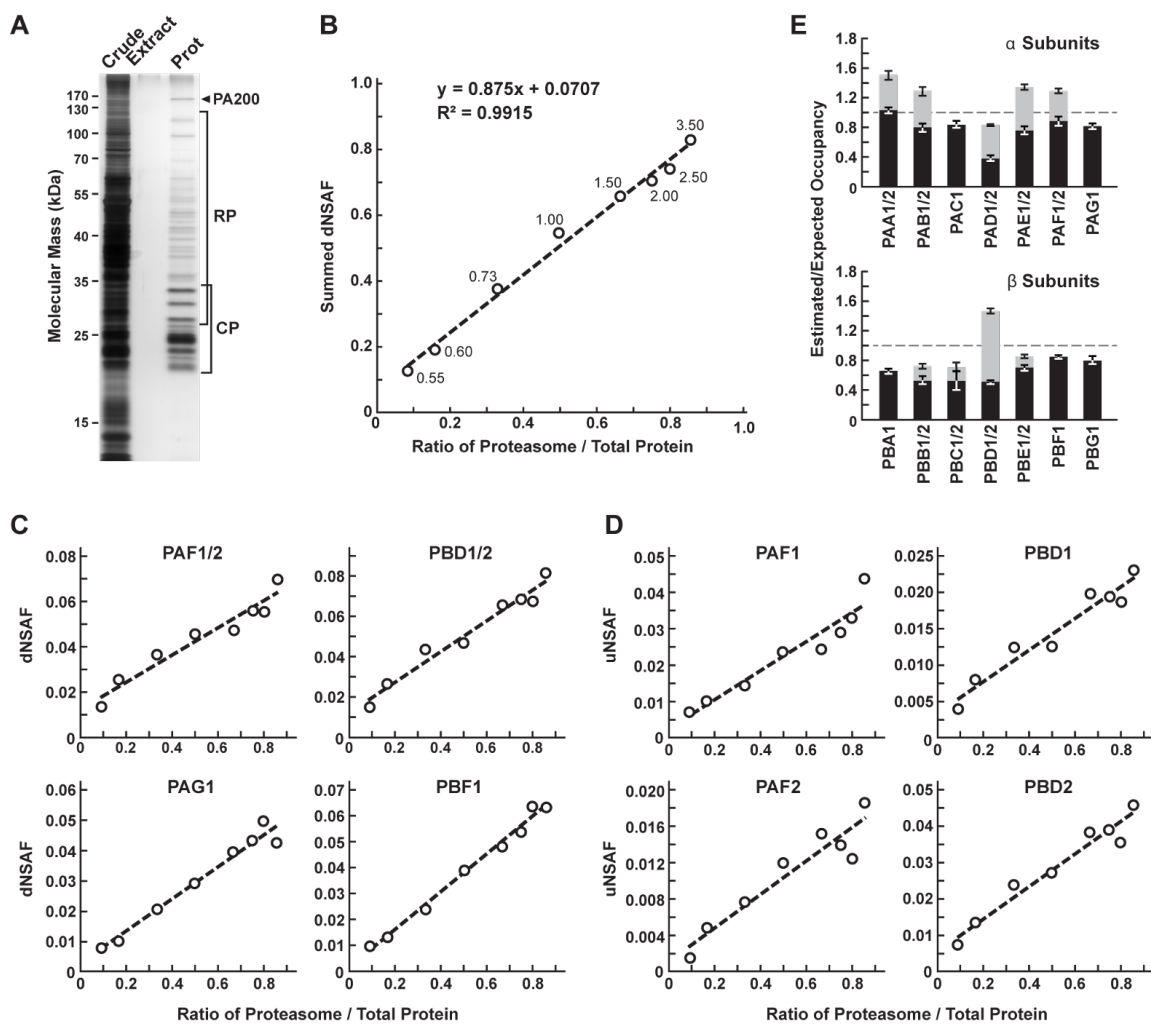


**Figure 2.3: Re- Analysis of MS/MS datasets generated with the Universal Proteome Standard 2 (UPS2).** The array of UPS2 standards were spiked into *Xenopus laevis* egg (Top) and embryo (Bottom) extracts at a range of concentrations. Following MS/MS analysis, dNSAF values for each protein were determined by Morpheus and MSpC with a change from Figure 2.2 in that two unique peptides were required to quantify a protein. (A) A log-log plot of dNSAF versus concentration for each UPS2 protein detected across each fmol range. (B) A log-log plot of average dNSAF vs average concentration of each group of UPS2 proteins at each fmol range: (50, 500, 5000, and 50,000 fmol).



Figure 2.3 A to Figure 2.2 A). Consequently, caution should be exercised when selecting this option even though it might provide a slight improvement in stringency (see supplemental discussion in Supporting Information). To demonstrate the utility and accuracy of MSpC as applied to our work, we analyzed 20S proteasomes isolated from *Arabidopsis thaliana*. This particle contains multiple subunits assembled in stoichiometric amounts, with many subunits encoded by two paralogous genes of sufficient amino acid identity (typically >90% (?)) such that discrimination between paralogs can be challenging using LFQ approaches (Book et al., 2010). To simulate changes in 20S proteasome abundance, we added varying amounts of trypsinized proteasomes (0.05  $\mu\text{g}$  to 3  $\mu\text{g}$ ) to a fixed amount of trypsinized *E. coli* lysate (0.5  $\mu\text{g}$ ) to generate proteasome/lysate ratios of ~0.091, 0.167, 0.333, 0.500, 0.667, 0.750, 0.800, 0.857. The digests were then subjected to MS/MS and the dNSAF value for each subunit along with the uNSAF value for individual isoforms were calculated by the Morpheus/MSpC pipeline (see Supplemental Methods). The data from this experiment are deposited in PRIDE with ID PXD003002. As shown in Figure 2.4, MSpC provided an excellent determination for the overall abundance of 20S proteasomes within a complex mixture, along with a good reflection of the abundance of individual subunits and their isoforms. When the dNSAF values for all subunits for the *Arabidopsis* 20S proteasome including their isoforms (represent-

[Figure 2.4 caption follows on next page]



**Table 2.1: Table of adj\_NSFAF values for each 20S proteasome subunit (equivalent to dNSAF) generated by analyzing the proteasome spike in experiments with the TPP and ABACUS pipeline. The top half of the table lists  $\alpha$ 1-7 (PAA-PAG) where 1 or 2 represent different isoforms) subunits, while the bottom half of the table lists  $\beta$ 1-7 subunits (PBA-PBG where 1 or 2 represent different isoforms)**

Ratio	0.09	0.17	0.33	0.50	0.67	0.75	0.80	0.86	Pearsons
PAA1	125.16	166.81	245.61	355.60	491.76	494.18	504.75	554.97	0.987788288683398
PAA2	57.52	70.52	114.60	126.35	132.06	162.25	358.13	154.62	0.493057505529942
PAB1	96.13	134.98	207.90	328.16	362.21	416.46	303.60	397.39	0.865439067273664
PAB2	29.67	79.42	153.34	282.91	365.22	299.47	310.19	299.93	0.848913155268619
PAC1	95.27	112.45	214.24	376.66	389.05	361.39	512.73	488.62	0.924694988613313
PAD1	62.58	76.16	106.03	196.03	179.37	182.99	164.04	188.71	0.79353639882071
PAD2	39.88	68.33	145.34	178.43	190.84	246.20	243.28	240.01	0.953695573961697
PAE1	50.41	97.53	193.18	368.87	464.79	397.27	468.64	534.16	0.950854292773049
PAE2	71.38	80.07	184.85	255.98	299.41	290.16	304.77	348.32	0.954954859737766
PAF1	132.67	179.60	207.68	253.51	253.04	327.11	334.73	460.76	0.834748054996771
PAF2	0.00	0.00	59.62	70.93	111.83	80.29	63.08	65.86	0.605990529937574
PAG1	108.09	142.82	271.18	389.94	495.94	530.90	632.72	540.86	0.96630247289725
Ratio	0.09	0.17	0.33	0.50	0.67	0.75	0.80	0.86	Pearsons
PBA1	85.99	113.01	208.38	233.44	273.50	327.23	417.12	560.62	0.850921399221317
PBB1	47.77	80.66	195.35	324.53	303.26	220.36	200.51	216.77	0.440413945248165
PBB2	17.11	4.54	0.00	0.00	21.05	102.51	144.30	187.86	0.618228756945012
PBC1	27.19	40.65	165.46	345.14	364.19	386.59	338.47	373.16	0.885192639364864
PBC2	10.64	40.65	0.00	55.58	145.43	158.90	159.74	228.05	0.844980229680132
PBD1	50.14	94.46	168.58	167.50	244.25	248.17	252.65	319.61	0.945107382530759
PBD2	81.45	149.28	256.88	319.89	390.60	407.02	380.30	490.17	0.950089780554562
PBE1	81.01	107.76	196.21	302.90	365.03	404.25	378.67	441.03	0.981079214776472
PBE2	0.00	1.52	5.59	8.05	40.06	56.56	59.39	58.56	0.886783231269529
PBF1	71.53	112.26	190.45	226.89	259.90	366.60	431.42	420.23	0.94257878462495
PBG1	104.93	174.55	232.34	392.37	425.10	397.85	396.42	463.13	0.905749458006341
									Average
									0.844830435248515

ing 14 distinct subunits, 10 of which exist as isoform pairs) were summed, a very close approximation of the dNSAF/actual abundance was obtained (slope=0.875) with a very strong linear correlation ( $R^2 = 0.99$ ) over a ~10-fold range in protein abundance. When each 20S proteasome subunit was analyzed individually, a strong linear response was also obtained ( $R^2 > 0.90$ ) for a majority of subunits (Figure 2.4 C and Supplemental Table 1).

For example, reasonably accurate concentration plots were obtained for the

---

**Figure 2.4 (preceding page): Confirmation of MSpC accuracy by analysis of MS/MS datasets generated with affinity purified Arabidopsis 20S proteasomes spiked into a total cell lysate from *E. coli*. Following MS/MS analysis, the dNSAF and uNSAF values for each subunit/isoform were determined by Morpheus and MSpC. (A)** A silver-stained SDS-PAGE gel of 20S proteasome samples affinity purified from 10-d-old Arabidopsis seedlings. The crude seedling extract (CE), sample buffer (SB), and affinity-purified 20S proteasome samples (Prot) are shown. **(B)** Quantification of trypsinized 20S proteasomes when mixed at varying ratios with trypsinized total protein lysates from *E. coli*. The spiked samples were subjected to MS/MS followed by data analysis with the Morpheus and MSpC. dNSAF values for each proteasome subunit were averaged across three technical replicates, then summed to obtain an estimate of abundance for the 20S proteasome, and plotted against their known ratios. The total protein load is listed at each point in ug. **(C and D)** dNSAF and uNSAF values determined from the data in panel B for individual subunits **(C)** and their isoforms **(D)** for several subunits of the 20S proteasome. **(E)** Quantification accuracy of the Morpheus/MSpC pipeline for determining the amount of each  $\alpha$  and  $\beta$  subunit of the 20S proteasome. Single subunit isoforms are in black, whereas subunits having two isoforms are shown in black and grey to reflect the contributions of isoforms 1 and 2 respectively. Each bar represents the average of eight technical replicates ( $\pm$  SE). The dashed line represents the expected value of one assuming an equal stoichiometry of each subunit within the particle.

PAF ( $\alpha 6$ ) and PBD ( $\beta 4$ ) subunits that are encoded by the PAF1/2 and PBF1/2 gene pairs, and for the PAG ( $\alpha 7$ ) and PBF ( $\beta 6$ ) subunits that are encoded by single PAG1 and PBF1 genes ( $R^2$  from 0.94 to 0.99). Even when we calculated uNSAF values for individual isoforms added to the *E. coli* lysate, strong linear responses were obtained (e.g., the PAF1/PAF2 and PBD1/PBD2 pairs) with robust correlations ( $R^2$  from 0.89 to 0.95) (Figure 2D). Taken together, MSpC worked well for relative LFQ analysis of a multi-subunit complex and its individual subunits and isoforms within a complex proteomic mixture.

The Morpheus/MSpC pipeline also allowed us to calculate the respective incorporation of each paralog in the complex (see Supplemental Methods). As shown in Figure 2E, these estimated/expected occupancies were close to unity for most subunits within both the  $\alpha$  and  $\beta$  rings of the 20S proteasome. The only strong deviation was for PBD1/2 ( $\beta 4$ ), which had a greater dNSAF value relative to other  $\beta$  subunits across the experiments analyzed (Supplemental Table 1). The calculations for uNSAF values also estimated the relative proportion of each isoform within the complex for those subunits expressed from paralogous genes. The data obtained are similar to prior studies of the complex involving quantitative top-down proteomic analysis of purified proteasome samples using ultra violet-intrinsic fluorescence to quantify tyrosine-containing subunits (Russell et al., 2013). However, our MSpC

analysis provided a more complete picture as several subunit isoforms were difficult to quantify by fluorescence either because they lacked tryosine, or because their fluorescence peaks overlapped with those of other subunits/isoforms. Notably, the protein isoform ratios measured here agree well with the expression ratios for the paralogous genes (Book et al., 2010), suggesting that the protein isoform abundance generally reflects the relative transcriptional activity of the gene pair. We consistently estimated slightly more  $\alpha$  ring subunits (PAA-PAG) versus  $\beta$  ring subunits (PBA-PBG) in the final MSpC calculations (Figure 2E). This deviation could represent enhanced detection of  $\alpha$  ring versus  $\beta$  ring subunits, or more likely that purification via the tagged  $\alpha$  ring subunit PAG1 also isolated assembly intermediates comprised of only  $\alpha$  ring subunits.

We compared the Morpheus and MSpC pipeline to the next most comparable open source, spectral-count-based LFQ pipeline, The Trans Proteomic Pipeline (TPP) (Deutsch et al., 2010) and ABACUS (Fermin et al., 2011) using our datasets generated with the 20S proteasome/*E. coli* lysate mixture (Supplemental Table 1 and 2). Morpheus/MSpC slightly outperformed TPP/ABACUS by having a greater overall accuracy (average linearity of 0.88 compared to 0.84), and by having more subunits showing an  $R^2$  linear correlation greater than 0.9 (14/23 subunits for MSpC versus 11/23 for ABACUS). In addition to this modest improvement, we note

that the Morpheus/MSpC pipeline required significantly less intermediary steps, thus accelerating the data analysis. Some of the additional steps in TPP/ABACUS could be automated from the command-line, but it would likely be a challenge for the average user. Importantly, we found that the Morpheus/MSpC pipeline was faster. Timing tests using the proteasome/*E. coli* spike data generated here showed that the Morpheus/MSpC pipeline was 1.9-fold faster than the TPP/ABACUS pipeline (Figure 3). Such an improvement was expected given that Morpheus completes its searches on average 1.3 to 4.6 times faster than most other search engines available (Wenger and Coon, 2013). Given its simplicity of use, speed, and open source nature, MSpC combined with Morpheus is clearly advantageous over other PSM-based LFQ approaches currently available. Moreover, by being open source, MSpC should allow others to extend its utility and to serve as a platform for integrating additional open source LFQ approaches into the Morpheus pipeline.

### 2.3. Methods

### 2.4. Tutorial

### 2.5. Acknowledgements

D.C.G. was supported by a grant from the U.S. Department of Energy Office of Science; Office of Basic Energy Sciences; Chemical Sciences, Geosciences, and Biosciences Division (DE-FG02-88ER13968) and a graduate training fellowship from the NIH (5 T32 GM 7133-37). M.S. and L.M.S were supported by a grant from the National Institute of Health/National Institute of General Medical Sciences (1P50HG004942). The authors thank Erin Gemperline, Richard S. Marshall, and Josh Coon for critical reading of the manuscript, and additionally thank Derek Bailey for a critical code review.

### 2.6. References

- Book, A.J., Gladman, N.P., Lee, S.S., Scalf, M., Smith, L.M., and Vierstra, R.D.** (2010). Affinity purification of the arabidopsis 26s proteasome reveals a diverse array of plant proteolytic complexes. *J. Biol. Chem.* **285**(33):25554–25569.
- Cox, J., Hein, M.Y., Lubner, C.A., Paron, I., Nagaraj, N., and Mann, M.** (2014). Accu-



rate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed maxlfq. *Mol. Cell. Proteomics* **13**(9):2513–2526.

**Deutsch, E.W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J.K., Martin, D.B., Nesvizhskii, A.I., and Aebersold, R.** (2010). A guided tour of the trans-proteomic pipeline. *Proteomics* **10**(6):1150–1159.

**Fermin, D., Basrur, V., Yocum, A.K., and Nesvizhskii, A.I.** (2011). Abacus: a computational tool for extracting and pre-processing spectral count data for label-free quantitative proteomic analysis. *Proteomics* **11**(7):1340–1345.

**Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W., and Gygi, S.P.** (2003). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem ms. *Proc. Natl. Acad. Sci. USA* **100**(12):6940–6945.

**Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M.** (2002). Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**(5):376–386.

- Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlett-Jones, M., He, F., Jacobson, A., and Pappin, D.J.** (2004). Multiplexed protein quantitation in *saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3**(12):1154–1169.
- Russell, J.D., Scalf, M., Book, A.J., Lador, D.T., Vierstra, R.D., Smith, L.M., and Coon, J.J.** (2013). Characterization and quantification of intact 26S proteasome proteins by real-time measurement of intrinsic fluorescence prior to top-down mass spectrometry. *PLoS One* **8**(3):e58157.
- Thompson, A., Schafer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A.K., and Hamon, C.** (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by ms/ms. *Anal. Chem.* **75**(8):1895–1904.
- Tu, C., Li, J., Sheng, Q., Zhang, M., and Qu, J.** (2014). Systematic assessment of survey scan and ms<sup>2</sup>-based abundance strategies for label-free quantitative proteomics using high-resolution ms data. *J. Proteome Res.* **13**(4):2069–2079.
- Wenger, C.D. and Coon, J.J.** (2013). A proteomics search algorithm specifically

designed for high-resolution tandem mass spectra. *J. Proteome Res.* **12**(3):1377–1386.

**Wong, J.W. and Cagney, G.** (2010). An overview of label-free quantitation methods in proteomics by mass spectrometry. *Methods Mol. Biol.* **604**:273–283.

**Zhang, B., VerBerkmoes, N.C., Langston, M.A., Uberbacher, E., Hettich, R.L., and Samatova, N.F.** (2006). Detecting differential and correlated protein expression in label-free shotgun proteomics. *J. Proteome Res.* **5**(11):2909–2918.

**Zhang, Y., Wen, Z., Washburn, M.P., and Florens, L.** (2010). Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Anal. Chem.* **82**(6):2272–2281.

**Zybailov, B., Mosley, A.L., Sardi, M.E., Coleman, M.K., Florens, L., and Washburn, M.P.** (2006). Statistical analysis of membrane proteome expression changes in *saccharomyces cerevisiae*. *J. Proteome Res.* **5**(9):2339–2347.

## COLOPHON

This document was typeset with  $\text{\LaTeX}$ . It is based on the University of Wisconsin dissertation template created by William C. Benton (available at <https://github.com/willb/wi-thesis-template>).